

---

## Research and Applications

# Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data

Na Hong, Andrew Wen, Feichen Shen, Sunghwan Sohn, Chen Wang, Hongfang Liu, and Guoqian Jiang

Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA

Corresponding Author: Guoqian Jiang, MD, PhD, Department of Health Sciences Research, Mayo Clinic, 200 First Street, SW, Rochester, MN 55905, USA; [jiang.guoqian@mayo.edu](mailto:jiang.guoqian@mayo.edu)

Received 25 June 2019; Revised 23 September 2019; Editorial Decision 25 September 2019; Accepted 1 October 2019

### ABSTRACT

**Objective:** To design, develop, and evaluate a scalable clinical data normalization pipeline for standardizing unstructured electronic health record (EHR) data leveraging the HL7 Fast Healthcare Interoperability Resources (FHIR) specification.

**Methods:** We established an FHIR-based clinical data normalization pipeline known as NLP2FHIR that mainly comprises: (1) a module for a core natural language processing (NLP) engine with an FHIR-based type system; (2) a module for integrating structured data; and (3) a module for content normalization. We evaluated the FHIR modeling capability focusing on core clinical resources such as Condition, Procedure, MedicationStatement (including Medication), and FamilyMemberHistory using Mayo Clinic's unstructured EHR data. We constructed a gold standard reusing annotation corpora from previous NLP projects.

**Results:** A total of 30 mapping rules, 62 normalization rules, and 11 NLP-specific FHIR extensions were created and implemented in the NLP2FHIR pipeline. The elements that need to integrate structured data from each clinical resource were identified. The performance of unstructured data modeling achieved *F* scores ranging from 0.69 to 0.99 for various FHIR element representations (0.69–0.99 for Condition; 0.75–0.84 for Procedure; 0.71–0.99 for MedicationStatement; and 0.75–0.95 for FamilyMemberHistory).

**Conclusion:** We demonstrated that the NLP2FHIR pipeline is feasible for modeling unstructured EHR data and integrating structured elements into the model. The outcomes of this work provide standards-based tools of clinical data normalization that is indispensable for enabling portable EHR-driven phenotyping and large-scale data analytics, as well as useful insights for future developments of the FHIR specifications with regard to handling unstructured clinical data.

**Key words:** data standards, electronic health records, Fast Healthcare Interoperability Resources, natural language process

---

## INTRODUCTION

With the widespread adoption of electronic health records (EHRs) in healthcare organizations, there is ample opportunity for secondary use of EHR data in clinical and translational research. However, the lack of EHR data interoperability between institutions makes it challenging to

integrate and share healthcare and clinical research data, thus impeding effective and efficient collaboration. A standardized model for data representation would assist in promoting the exchange of EHR data, achieving large-scale data-driven research collaborations and supporting rapid generation of accurate and computable phenotypes.

As the next generation standards framework, the Fast Healthcare Interoperability Resources (FHIR)<sup>1</sup> was developed by HL7 to meet clinical interoperability needs. FHIR defines a collection of “resources” that “can easily be assembled into working systems that solve real-world clinical and administrative problems at a fraction of the price of existing alternatives.” This assembly process typically requires “profiling”—the adaptation of the FHIR core resources for use in particular contexts and use cases. FHIR also leverages the latest web standards and places a strong focus on implementability. Notably, major EHR vendors (eg, Epic, Cerner) and healthcare providers (eg, Mayo Clinic, Intermountain Healthcare, and Partners Healthcare) have been involved in the development and adoption of FHIR through HL7 Argonaut Project.<sup>2</sup> In the clinical research domain, the FHIR-based standard Application Programming Interfaces (APIs) have been leveraged in a national collaboration known as the Sync for Science (S4S) initiative<sup>3</sup> to help patient share EHR data with researchers and empower individuals to participate in health research.

While FHIR is rapidly being adopted in different EHR systems at various institutions, there are a number of gaps on how to represent unstructured information in clinical narratives using FHIR. First, there are unmet needs on standardizing unstructured clinical data. The recent proposal from the Office of the National Coordinator for Health Information Technology (ONC) and the Centers for Medicare & Medicaid Services (CMS) that FHIR APIs be required for certified EHR systems<sup>4</sup> highlighted the importance of the FHIR standard. Particularly, using NLP to gain access to the narrative content in EHRs via FHIR will be of great value to data analytics, quality improvement, and advanced decision support. However, current HL7 Argonaut project has not yet provided a solution to standardize unstructured data. Second, although it is certainly feasible to use the FHIR Composition<sup>5</sup> as a document resource for representing clinical narratives in EHRs, few studies have been done on (1) the tool development for generating the FHIR resource instances from clinical narratives leveraging the NLP technology; and (2) assessing the discrepancies between FHIR data models and NLP type systems. The DeepPhe project<sup>6</sup> created a conceptual model (DeepPhe Ontology) leveraging FHIR models to provide a terminology of entities and relationships between them to represent cancer phenotypes extracted from unstructured EHR data.<sup>7</sup> The DeepPhe project was focused more on adapting FHIR data models to represent cancer phenotypes, rather than developing a data normalization pipeline to formally model unstructured data and NLP outputs using FHIR specification. In addition, no work has been done to assess whether FHIR can represent core elements (eg, negation, certainty, etc.) from different clinical NLP systems for handling unstructured clinical data. Third, a common type system for clinical NLP has been regarded as an important way to enable interoperability between structured and unstructured data generated in different clinical settings.<sup>8</sup> As a part of SHARPN data normalization pipeline, cTAKES<sup>9</sup> implemented a common type system that has an end target of deep semantics based on the clinical element models (CEMs).<sup>10</sup> In the context of secondary use of EHR data, we envision that an FHIR standard-based common type system would better improve semantic interoperability between heterogeneous clinical data sources, given the rapid adoption of FHIR as an international standard in different EHR systems. This novel FHIR-based type system not only can enable effective exchange, integration, sharing, and reuse of encoded and structured clinical narratives, along with well-structured EHR data, but it can also serve as target data models for advanced development of NLP system. The latter includes the following two inno-

vative aspects: (1) a well-defined target data model based on the FHIR type system allows us to easily integrate multiple distinct NLP pipelines, each of which may have their own specialties; and (2) FHIR provides a powerful modeling mechanism that enables the creation of new standard models for particular NLP-based information retrieval tasks, for example, cancer-specific phenotype extraction.

The objective of this study was to design, develop, and evaluate a scalable and standards-based EHR data modeling framework and accompanying clinical data normalization pipeline leveraging the HL7 FHIR specification. We implemented a generic pipeline known as NLP2FHIR for modeling unstructured EHR data using the FHIR specification and evaluated the main outcomes as well as the performance of our pipeline using the EHR data from the Mayo Clinic.

## MATERIALS AND METHODS

### Materials

#### Clinical narrative corpora

To support the experiment and evaluation of the NLP2FHIR pipeline, a FHIR-based clinical data normalization pipeline, we reused a corpus of 734 clinical notes from Mayo Clinic’s previous clinical NLP research projects, including SHARPN, the Mayo MedXN project, and the Mayo Clinic’s family member history (FMH) NLP project.<sup>11–13</sup> These notes were randomly collected from Mayo Clinic’s EHR. Four section types (ie, problem list, family history, medication list, and past procedure list) with 940 individual sections were used for the unstructured data modeling study in this study. These corpora had previously been annotated by clinical subject matter experts for research purposes.

#### UIMA-based clinical NLP tools

UIMA, short for Unstructured Information Management Architecture, is a data-driven architecture where individual components are able to communicate with one another through a data structure called the common analysis system (CAS), which uses a specified hierarchical type system. The type system allows for flexible passing of input and output data types between components of an NLP system. In this study, the NLP2FHIR pipeline implementation integrated three UIMA-based clinical NLP tools as follows: (1) cTAKES,<sup>9</sup> an open-source NLP system for extraction of information from EHR clinical free-text, which provides a tool for selecting different descriptors to support common clinical NLP tasks (eg, part-of-speech tagging, chunking, and dictionary lookup); (2) MedXN,<sup>12</sup> an open-source medication entity/attribute extraction and normalization tool, which extracts comprehensive medication information and normalize it to the most appropriate RxNorm concept unique identifier (RxCUI) as specifically as possible; and (3) MedTime,<sup>14</sup> an open-source temporal information detection system, which extracts EVENT/TIMEX3 and temporal link (TLINK) identification from clinical text.

#### FHIR specification and application programming interface

The building block in FHIR is a Resource,<sup>1</sup> which provides a common way to define and represent all exchangeable content and related metadata in a particular modeling domain. In this study, we leveraged both document resources Composition/Bundle and clinical resources Condition, Procedure, MedicationStatement/Medication, and FamilyMemberHistory to model unstructured EHR data and NLP outputs. As of September 10, 2019, the version FHIR R4 has been released officially while we used an earlier version, the

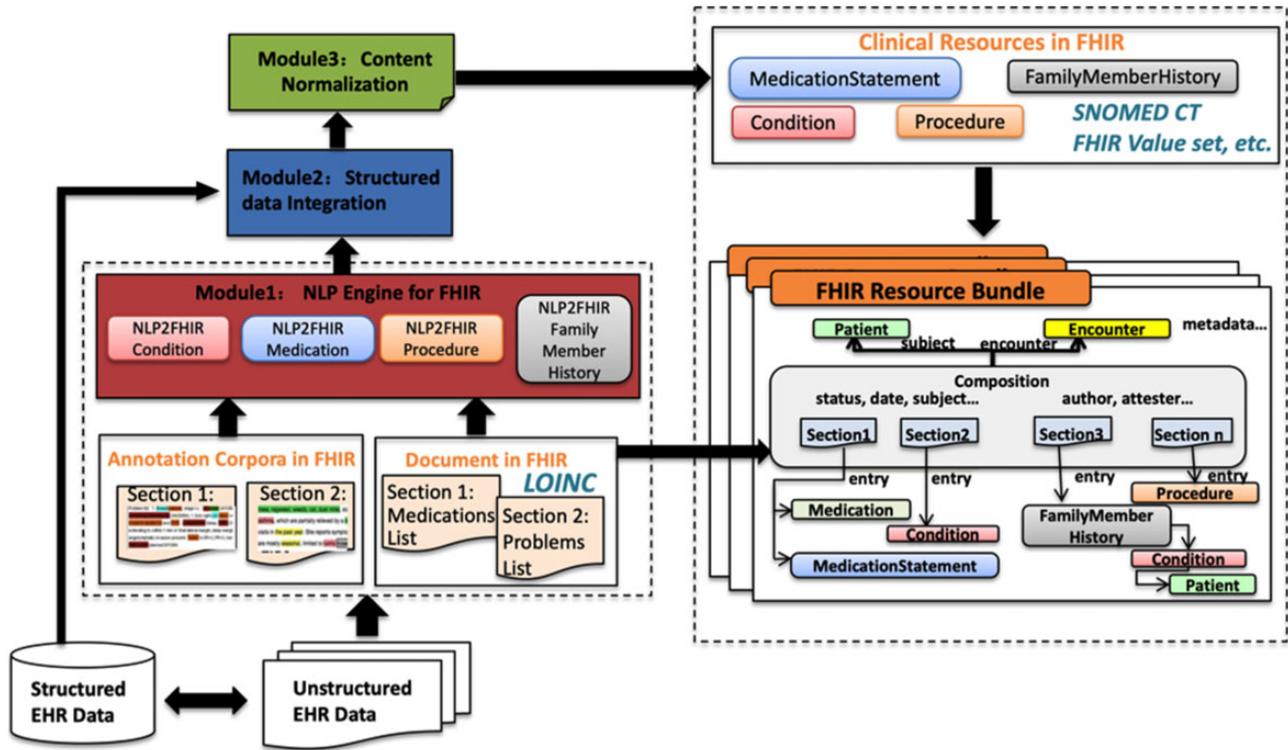


Figure 1. NLP2FHIR pipeline for EHR data modeling. EHR: electronic health record; FHIR: Fast Healthcare Interoperability Resources.

Standards for Trial Use Version 3 (STU3) (released on April 19, 2017) in this study.

As the FHIR modeling interface, HAPI FHIR was used in our implementation to support FHIR data modeling. HAPI FHIR is an open-source implementation of the FHIR specification in Java.<sup>15</sup> HAPI FHIR defines model classes for every resource type and data type defined by FHIR specification. In addition, HAPI supports data validation within FHIR modeling. Therefore, we used the HAPI FHIR application programming interface to serialize elements extracted from clinical documents into standard FHIR eXtensible Markup Language (XML) and JavaScript Object Notation (JSON) representations.

Methods

Figure 1 shows the system architecture of the FHIR-based clinical data normalization pipeline. The NLP2FHIR pipeline comprises the following three modules: (1) a module for a core NLP engine with an FHIR-based type system, (2) a module for integrating structured data, and (3) a module for content normalization. In addition, an intuitive graphical user interface is implemented to allow users to configure the pipeline with parameters in terms of unified medical language system (UMLS) username and password, input directory and type (eg, TEXT, or COMPOSITION\_RESOURCE), section definition directory and file, resources to produce, and output directory and formats (eg, FHIR JSON). Figure 2 shows a screenshot of the graphic user interface of the implemented NLP2FHIR pipeline.

Module for a core NLP engine

Unstructured clinical documents (eg, clinical notes, radiology reports) usually convey large amounts of valuable information. The module for modeling unstructured data contains the following components:

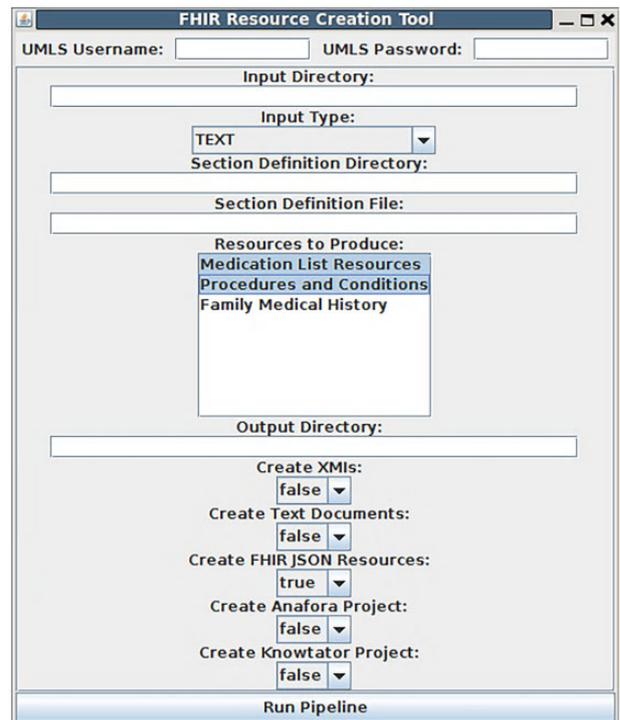


Figure 2. A screenshot of the graphic user interface of the implemented NLP2FHIR pipeline. Input type can be COMPOSITION\_RESOURCE, BUNDLE\_RESOURCE, XMI, or TEXT.

1. Rendering clinical documents in FHIR Composition resource as input: As a type of FHIR document resource, the Composition resource<sup>5</sup> defines a set of elements that are assembled together

into a single logical document and provides a coherent statement for meaningful document representation. Therefore, we used the FHIR Composition resource to standardize variants of clinical documents as standard input of the FHIR NLP engine. A collection of standard Logical Observation Identifiers and Codes (LOINC) codes were assigned for encoding sections, including Reported Problem List (11450-4), History of Present Illness Narrative (10164-2), History of Medication Use Narrative (10160-0), History of Procedures Document (47519-4), and History of Family Member Diseases Narrative (10157-6). After analyzing section content and FHIR resource definition, we created mappings from the sections to FHIR resources.

2. *Integrating existing clinical NLP tools as the NLP engine of the NLP2FHIR pipeline:* We integrated the existing clinical NLP tools, comprising cTAKES, MedXN, and MedTime to extract clinical entities from corresponding document sections, and standardized them using the FHIR resources Condition, Procedure, MedicationStatement (including Medication), and FamilyMemberHistory. Different tools were used to handle different clinical narrative extraction tasks. The cTAKES and MedTime were used for the FHIR element entity and relation extraction tasks from the problem list (corresponding to FHIR Condition), past history of surgery (corresponding to FHIR Procedure), and FMH (corresponding to FHIR FamilyMemberHistory). For medication list (corresponding to the FHIR MedicationStatement and Medication), MedXN and MedTime were set up for extracting and standardizing drug names and drug related temporal expressions.
3. *Creating a FHIR-based type system to interoperate with UIMA-based NLP tools:* UIMA provides a software framework for building type systems while supporting interaction between multiple NLP components. To allow for rapid integration of the NLP tooling output or particular FHIR element extraction results, we generated an FHIR-based type system using the FHIR Standards for Trial Use (STU) 3 v1.8.0 specification. The FHIR-based type system is used to meet the need of interoperability between different NLP pipelines, which enhances the NLP component interoperability through maintaining consistent naming of elements, structure hierarchy, and data restrictions present within the FHIR definitions.<sup>16</sup>
4. *Defining mapping rules:* We compared the NLP output types with FHIR specification, and reassembled extraction outputs of the NLP tools by creating mapping rules between heterogeneous NLP outputs and standard FHIR elements. FHIR resource and element mapping levels were conducted in terms of granularity at two different levels: (1) narrative sections to FHIR resources and (2) NLP output types to FHIR elements. In total, 30 different types and levels of mapping rules were generated to support integrating heterogeneous NLP outputs to our NLP2FHIR pipelines, and 59 target FHIR elements could be directly populated from NLP tools. [Table 1](#) shows examples of the mapping rules. Additional details are provided in Supplemental Material S2.
5. *Creating NLP-specific FHIR extensions:* We noticed that the current FHIR resource definition did not cover all the elements from NLP outputs, and some NLP-specific elements of these outputs were essential within the context of subsequent downstream analytics. Therefore, we created a list of FHIR extensions to keep these NLP-specific elements by analyzing a set of clinical NLP elements defined in the latest OHDSI CDM v6.0,<sup>17</sup> cTAKES-type system definitions, and input from NLP experts. [Table 2](#) shows a group of 11 common NLP-specific FHIR

extensions created for supporting extended unstructured EHR data normalization. The extension elements were aligned for semantic overlap or similarity by the NLP expert-based reviews. The NLP-specific elements defined in the FHIR extensions were reviewed using the following two basic criteria: (1) whether the element was commonly identified in clinical narratives; and (2) whether the existing NLP tools could handle the entity/relation element extraction.

#### Module for integrating structured data

Although the entity mentions that were extracted from clinical narratives using NLP tools covered the majority of the elements as defined in the FHIR Composition resource and its referenced clinical resources, there are still, however, several pieces of information that needed to be captured from structured EHR data and integrated with the NLP output to complete the population of the corresponding FHIR resource content. The crucial steps for integrating structured data with NLP output consisted of: (1) setting templates for mapping the structured source data elements to the corresponding FHIR resource elements; (2) extracting the instance data from the EHR, where normalization processing may have applied; (3) linking structured instance data with NLP output through a primary key reference (eg, patient id) or directly as an attribute defined within an FHIR resource. For example, when populating each instance of the FHIR MedicationStatement resource, we could directly get its subject information (ie, who is/was taking the medication) from structured EHR data and link each subject to the specific MedicationStatement instance through the Reference (Patient-Group) identifier. [Table 3](#) lists the information that was captured from structured EHR data and integrated with each component of the NLP2FHIR pipeline.

#### Module for content normalization

Content normalization makes the resource content conform to the FHIR specification in terms of its datatype definitions for corresponding model elements and its content semantics through terminology binding. As mentioned previously, we leveraged a number of core FHIR resources Condition, Procedure, MedicationStatement, and FamilyMemberHistory to capture clinical concepts identified from the unstructured narratives. Therefore, we followed the recommendation from the definitions of these core resources on the use of preferred code systems. In addition, many FHIR elements have specific datatype requirements, (eg, boolean, integer, string, and decimal), thus, we implemented the datatype conversion and value transformation to their target element definition. Handling terminology binding is one of the concept normalization tasks, which requires binding an FHIR element with the identity and version of a terminology system, the codes, and their display names, as shown in [Supplemental Table 2](#).

In addition to standard codes defined in external terminologies, FHIR also defines its own value sets with a list of codes in its specification. We created a set of transformation rules to normalize the element instances in terms of terminology binding. For example, tab is an instance for the element Medication.form, which is normalized to Tablet (385055001) defined in the SNOMED CT Form Codes. A number of NLP tools support the concept normalization for the identified entities. For instance, MedXN normalizes a variety of nonstandard medication mentions to the RxNorm codes, and cTAKES assigns UMLS concept unique identifiers to the extracted entities. However, the code systems recommended by FHIR may not

**Table 1.** Examples of mapping rules between EHR sources, NLP output types, and FHIR elements

Source	NLP output types	FHIR elements	Mapping types	Examples	
Medication list	MedXN: Drug	MedicationStatement.medicationCodable-Concept	1:1	Oxamniquine→[SNOMED: 747006]	
	MedXN: Drug: attributes: type="frequency"	MedicationStatement.dosage.timing.frequency MedicationStatement.dosage.timing.frequencyMax	1:n	Once daily →1[frequency], 1[period], d[periodUnit] 4–6 times →4[frequency], 6[frequencyMax]	
	MedTime: MedTimex3: type="SET"	MedicationStatement.dosage.timing.period MedicationStatement.dosage.timing.periodMax MedicationStatement.dosage.timing.periodUnit MedicationStatement.dosage.asNeeded.asNeededBoolean MedicationStatement.dosage.timing.dayofWeek MedicationStatement.dosage.timing.when		As needed for heel pain→true Regular: once daily every six hours Irregular: as needed for pain every Monday, Tuesday Wednesday	
	MedXN: Drug: attributes: type="duration"	Dosage.duration Dosage.durationMax Dosage.durationUnit	1:n	3 days →3[duration], d[durationUnit]	
	MedTime: MedTimex3: type="DURATION"				
	MedXN: Drug: attributes: type="route"	Dosage.route	1:1	By mouth [oral route]	
	MedXN: Drug: attributes: type="strength"	Medication.ingredient.amount.numerator.quantity.value Medication.ingredient.amount.numerator.quantity.unit Medication.ingredient.amount.denominator.quantity.value Medication.ingredient.amount.denominator.quantity.unit	1:n	Regular: 500 mg /5 mL→ 500[numerator.quantity.value], mg[numerator.quantity.unit], 5[denominator.quantity.value], mL [denominator.quantity.unit] Irregular: 200 mg → Default assign: 1[denominator.quantity.value]	
	MedXN: Drug: attributes: type="form"	Medication.form	1:1	tab[Tablet]	
	MedXN: Drug: attributes: type="dosage"	Dosage.doseQuantity.value Dosage.doseQuantity.unit Dosage.doseQuantity.Range.low.value Dosage.doseQuantity.Range.low.unit Dosage.doseQuantity.Range.high.value Dosage.doseQuantity.Range.high.unit	1:n	10 mL →10[value], mL[unit] 2–3 tabs →2[range.low.value], tab[range.low.unit], 3[range.high.value], tab[range.high.unit]	
	Problem list	cTAKES: Disease_disorder	Condition.code	1:1	The Lingering sore throat → [SNOMED: 140004] /Chronic pharyngitis
		cTAKES: Anatomical_Site relations: type="LocationOf"	Condition.bodySite	1:1	Back of the head → [SNOMED: 774007] / Head and neck
		cTAKES: modifier: type="Severity"	Condition.severity	1:1	Very bad → [SNOMED: 24484000] / Severe
	Family history	Relation relations: type:"SideOfFamily" relations: type:"Blood" relations: type:"Adopted"	FamilyMemberHistory.relationship	n:1	Grandpa→ MGRFTH / maternal grandfather
Laboratory test	test_code	Observation.code	1:1	Albumin in Semen →[LOINC: 10558-5] / Albumin [Moles/volume] in Semen	
Section	Source Section	Composition.section.code	1:1	Family history→ [LOINC: 10157-6] / History of family member diseases narrative	

(continued)

**Table 1.** continued

Source	NLP output types	FHIR elements	Mapping types	Examples
Temporal information	MedTime: MedTimex3: type- "DATE"	MedicationStatement.effectiveDatetime	1:1	April 16th
	MedTime: MedTimex3: type- "TIME"	MedicationStatement.effectiveDatetimeDo- sage.timeOfDay	1:n	April 8, 2008 at 04: 38 PM

**Table 2.** Proposed FHIR NLP extensions for clinical NLP

Proposed FHIR NLP extension	FHIR resource	Definition reference sources <sup>a</sup>	Description
offset	Any	[Ref: cTAKES/ LineAndTokenPosition] [Ref: OHDSI NLP/offset]	Token line and offset of the extracted term in the input note
raw_text	Any	[Ref: OHDSI NLP/ lexical_variant]	Raw text extracted from the NLP tool
context	Any	[Ref: cTAKES /LookupWindowAnnotation] [Ref: cTAKES /ContextAnnotation] [Ref: OHDSI NLP/snippet]	Contextual information of an entity
nlp_system	Any	[Ref: OHDSI NLP/nlp_system]	Name and version of the NLP system that extracted the term. Useful for data provenance
nlp_date/nlp_datetime	Any	[Ref: OHDSI NLP/nlp_date, nlp_datetime]	The date or datetime of the note processing. Useful for data provenance
term_temporal	Any	[Ref: cTAKES/HistoryOfModifier] [Ref: OHDSI NLP/term_temporal]	The time modifier associated with the extracted term
confidence_score	Any	NLP experts inputs	The confidence score indicates the probability of accuracy with the extracted term
conditional_modifier	Any	[Ref: cTAKES/ConditionalModifier]	Used to indicate that a procedure or assertion occurs under certain conditions
negated_modifier	Condition Procedure Medication	[Ref: cTAKES/ PolarityModifier]	Used to indicate that a procedure or assertion did not occur or does not exist
certainty_modifier	Condition	[Ref: cTAKES/ UncertaintyModifier]	An introduction of a measure of doubt into a statement
LabDeltaFlag_modifier	Observation	[Ref: cTAKES/ ssLabDeltaFlagModifier]	An indicator to warn that the laboratory test result has changed significantly from the previous identical laboratory test result

*Abbreviations:* FHIR: Fast Healthcare Interoperability Resources; NLP: natural language processing.

<sup>a</sup>For expansions of abbreviations used in definition reference sources, please refer to text.

be the same as those used in existing NLP tools. For example, FHIR suggests the use of SNOMED CT codes for the element “MedicationStatement.medicationCodeableConcept,” but we acquired its corresponding RxNorm codes from MedXN. For this situation, terminology mapping is necessary. Therefore, we used manually created transformation rules and leveraged existing terminology mappings as the main methods for content normalization. Although varieties of individual resources are produced by the standard outputs of our normalization pipeline, these resources are actually directly or indirectly relevant to each other.

According to the FHIR specification, we normalized various expressions from source EHR data using a group of normalization rules. A total of 62 normalization rules were created and implemented (Table 4). Other value set and data type conformations for each FHIR element are included in Supplementary Material S2.

In the FHIR specification, the Bundle resource<sup>18</sup> refers to a container for a collection of resources, which is typically used to gather a collection of resources into a single Bundle instance with a specific context. In this study, the FHIR Bundle resource is used to contain both the instances of the FHIR Composition resource and its referenced clinical resources. We developed a wrapping process as a part

of the NLP2FHIR pipeline for connecting individual resources into an exchangeable Bundle resource that preserves complete semantics to support secondary use of the standardized instance data. An example of the FHIR Bundle recourse is shown in Figure 3.

**Evaluation design**

The main purpose of the performance evaluation is to demonstrate whether the standardization process causes a loss in performance, as there are often concerns that standardization is culpable for the loss in performance due to data elements that are originally output by NLP being not representable in a standard (eg, word-sense disambiguation, bag-of-word ngrams, cooccurrences, etc.). The performance evaluation was conducted through the following steps.

1. *Reusing annotation corpora:* We reused the corpora from SHARPN, MedXN, and FMH projects from Mayo’s unstructured EHR data. These corpora contain annotations made by medical experts, the quality of which has been sufficiently verified through previous studies.
2. *Standardizing the annotation corpora using FHIR-based annotation schema:* To support corpora reuse and integration, we

**Table 3.** Structured data integrated from EHRs for the NLP2FHIR pipeline

NLP2FHIR pipeline	Elements populated from structured data	Data type	Definitions
Condition	Condition.clinicalStatus	CodeableConcept	active   recurrence   relapse   inactive   remission   resolved(HL7 ValueSet: ConditionClinicalStatusCodes)
	Condition.category	CodeableConcept	problem-list-item   encounter-diagnosis(HL7 ValueSet: ConditionCategoryCodes)
	Condition.subject	Reference	Who has the condition
	Condition.encounter	Reference	The encounter during which this condition was created or diagnosed
Procedure	Condition.recordedDate	dateTime	Date record was first recorded
	Procedure.status	code	A code specifying the state of the procedure. Generally, this will be the in-progress or completed state
	Procedure.subject	Reference	The person, animal or group on which the procedure was performed
	Procedure.category	CodeableConcept	Classification of the procedure
MedicationStatement	Procedure.encounter	Reference	The Encounter during which this Procedure was created or performed or to which the creation of the record is tightly associated
	MedicationStatement.status	code	active   completed   entered-in-error   intended   stopped   on-hold   unknown   not-taken
	MedicationStatement.subject	Reference	Who is/was taking the medication
FMH	MedicationStatement.category	CodeableConcept	Type of medication usage(SNOMED CT)
	MedicationStatement.dateAsserted	dateTime	When the statement was asserted
	FamilyMemberHistory.status	code	partial   completed   entered-in-error   health-unknown(HL7 ValueSet: FamilyHistoryStatus)
	FamilyMemberHistory.dataAbsentReason	CodeableConcept	subject-unknown   withheld   unable-to-obtain   deferred (HL7 ValueSet: FamilyHistoryAbsentReason)
	FamilyMemberHistory.patient	Reference	Patient history is about
	FamilyMemberHistory.date	dateTime	When history was recorded or last updated

*Abbreviations:* EHR: electronic health record; FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level Seven International; NLP: natural language processing.

**Table 4.** Normalization results for each NLP2FHIR pipeline

NLP2FHIR pipeline	No. of rules	Element examples	Data type	Normalization examples
MedicationStatement	25	MedicationStatement.medicationCodableConcept	CodeableConcept	Oxamniquine → 747006[coding.code]
		MedicationStatement.dosage.timing.frequency	integer	Once daily → 1[frequency]
		MedicationStatement.dosage.asNeeded.asNeededBoolean	boolean	As needed for heel pain → true
		MedicationStatement.dosage.timing.dayofWeek	code	Every Monday → mon[http://hl7.org/fhir/ValueSet/days-of-week]
Procedure	10	Procedure.code	CodeableConcept	Kidney echography → 306005/echography of kidney
		Procedure.reasonCode Procedure.performed[x].performedDateTime	CodeableConcept dateTime	134006/decreased hair growth April 16th, 2010
Condition	13	Condition.code	CodeableConcept	The Lingering sore throat → 140004/Chronic pharyngitis
		Condition.bodySite	CodeableConcept	774007/Head and neck
FMH	14	Condition.abatementString	string	Resolved
		FamilyMemberHistory.condition.code FamilyMemberHistory.relationship	CodeableConcept CodeableConcept	3511005/Infectious thyroiditis MGRFTH/maternal grandfather

*Abbreviations:* FHIR: Fast Healthcare Interoperability Resources; MGRFTH: a role code for maternal grandfather; NLP: natural language processing.

designed and developed a framework,<sup>19</sup> which contains the following two components: (1) an automatic schema transformation component, in which the annotation schema in each corpus is automatically transformed into the FHIR-based schema; and

(2) an expert-based verification and annotation component, in which existing annotations can be verified and new annotations can be added for new elements defined in FHIR. Three co-authors (NH, AW, and GJ) reviewed and verified the

```

{
  "resourceType": "Bundle",
  "id": "999c5b02-3070-4785-b43b-8dcb92fd92b8",
  "type": "document",
  "entry": [
    {
      "fullUrl": "Composition/ON03FP00001000235.txt",
      "resource": {
        "resourceType": "Composition",
        "id": "ON03FP00001000235.txt",
        "identifier": {
          "system": "urn:ietf:rfc:3986",
          "value": "urn:uuid:f3fe1c0a-39cf-3c8a-a9ed-cd053d814b76"
        }
      },
      "section": [
        {
          "code": {
            "coding": [
              {
                "system": "http://hl7.org/fhir/ValueSet/doc-
section-codes",
                "code": "11450-4"
              }
            ]
          },
          "text": {
            "status": "additional",
            "div": "<div xmlns='http://www.w3.org/1999/xhtml
\\>Problem list: 1. Breast cancer, stage I-c -CT chest (4/15/64)
right apical thickening, repeat CT 6mos. -Abnormal screening
mammogram (04/28/64): 1.3cm right LIQ mass with adjacent 6mm mass -R
core needle bx (5/04/64): invasive ductal CA and DCIS -Lumpectomy
&amp; SNbx (5/24/64): 1.4cm mass, grade 2/3, 5mm focus of DCIS
extending to within 1 mm of final lateral margin, deep margin of the
14mm mass positive for intraductal Ca, focal angiolymphatic invasion
present. Tumor is ER+3, PR+3, her2neu not overexpressed, Ki67&amp;lt;
10%. 0/4 SN negative. Reexcision planned 6/15/64</div>"
          },
          "entry": [
            {
              "reference": "Condition/edac5dae-423f-334b-b806-
a3c5acdfe9f"
            },
            {
              "reference": "Procedure/43f61928-5307-3a19-
a824-8ce446ea77d6"
            }
          ]
        }
      ]
    },
    {
      "fullUrl": "Condition/edac5dae-423f-334b-b806-a3c5acdfe9f",
      "resource": {
        "resourceType": "Condition",
        "id": "edac5dae-423f-334b-b806-a3c5acdfe9f",
        "identifier": {
          "system": "urn:ietf:rfc:3986",
          "value": "urn:uuid:edac5dae-423f-334b-b806-a3c5acdfe9f"
        }
      },
      "code": {
        "coding": [
          {
            "system": "http://snomed.info/sct",
            "code": "190121004"
          }
        ]
      },
      "text": "Breast cancer"
    },
    {
      "bodySite": [
        {
          "coding": [
            {
              "system": "http://snomed.info/sct",
              "code": "361079003"
            }
          ]
        },
        {
          "text": "Breast"
        }
      ],
      "abatementString": "positive"
    }
  ]
}

```

**Figure 3.** Example of the FHIR bundle resource with a standard section of “Problem List—Reported (LOINC: 11450-4)” and its referenced FHIR resources. FHIR: Fast Healthcare Interoperability Resources; LOINC: Logical Observation Identifiers and Codes.

annotations. NH and AW have extensive experience in medical informatics, FHIR-based research applications, and clinical NLPs; and GJ has medical background with extensive expertise in medical informatics and standards-based research applications. The generated FHIR-represented corpora were used as the gold standard to facilitate the FHIR NLP engine performance tuning and evaluation.

3. *Evaluating the performance of the NLP2FHIR pipeline:* We used standard measures (precision, recall, and *F* score) using the FHIR-based annotation corpora as the gold standard. Based on the NLP output mapping and machine learning methods integration, we evaluated the core element extraction and normalization performance of the FHIR resources Condition, Procedure, MedicationStatement, and FamilyMemberHistory. As the FHIR model contained more granular clinical elements than those output types from existing NLP tools, our FHIR NLP engine also supported the particular FHIR element extraction algorithms leveraging machine learning methods; three annotation corpora were used for different FHIR element machine learning tasks.

## RESULTS

We measured the performance of our pipeline that achieved *F*-scores ranging from 0.690 to 0.995 for various FHIR element representations, which is comparable to the general clinical NLP tasks.<sup>9,12,14</sup>

The performance results of core elements and original baseline tools are shown in Table 5.

The results demonstrated that the NLP2FHIR pipeline does not cause a decrease in performance through our integration framework, which was established to enhance EHR interoperability compared with diverse existing tools. The element FamilyMemberHistory.extension.negated\_modifier is one of the FHIR NLP-specific extensions, and its performance results were based on the cTAKES outputs; the element FamilyMemberHistory.relationship was newly identified using the machine learning-based relation extraction algorithm, and other element evaluation was based on mappings and normalization rules for existing NLP tools. Therefore, the results verified the feasibility of the NLP2FHIR pipeline on standardizing unstructured EHR data.

## DISCUSSION

The use of standards in modeling EHR data has the potential to unlock clinical data interoperation, high-throughput computation, and meaningful use.<sup>20</sup> To promote FHIR-oriented EHR data modeling, we designed and developed a FHIR-based clinical data normalization pipeline (ie, NLP2FHIR) that can extract, standardize, integrated data from unstructured clinical narratives. We believe that modeling unstructured EHR data using the NLP2FHIR pipeline can play an important role in enabling advanced semantic interoperability across different EHR systems.

**Table 5.** Evaluation results on the performance of the NLP2FHIR pipeline

FHIR resource	FHIR element	Precision	Recall	F score	Baseline F Score
MedicationStatement and Medication	MedicationStatement.medicationCodeableConcept	0.996	0.982	0.988	MedXN: 0.581–0.954
	Dosage.timing.repeat.frequency	0.795	0.873	0.832	MedTime:
	Dosage.timing.repeat.period	0.959	0.914	0.936	0.880
	Dosage.timing.repeat.duration	0.600	1	0.750	
	Dosage.route	0.957	0.816	0.878	
	Medication.ingredient.amount.numerator.quantity.value	0.930	0.815	0.869	
	Medication.ingredient.amount.numerator.quantity.unit	0.926	0.899	0.911	
	Medication.form	0.871	0.704	0.779	
	Dosage.timing.repeat.when	1	0.571	0.727	
Condition	Dosage.asNeededBoolean	0.913	0.583	0.712	
	Condition.code	0.865	0.696	0.771	cTAKES:
	Condition.bodySite	0.871	0.611	0.718	0.768–0.954
	Condition.severity	0.909	0.556	0.690	
Procedure	Condition.extension.negated_modifier	0.992	0.998	0.995	
	Procedure.code	0.889	0.643	0.746	cTAKES:
	Procedure.bodySite	0.895	0.798	0.844	0.768–0.954
FamilyMemberHistory	FamilyMemberHistory.condition.code	0.940	0.716	0.813	cTAKES:
	FamilyMemberHistory.extension.negated_modifier	0.937	0.967	0.952	0.768–0.954
	FamilyMemberHistory.relationship	0.756	0.739	0.747	

*Abbreviations:* FHIR: Fast Healthcare Interoperability Resources; NLP: natural language processing.

The key contributions of our study are: (1) the creation of mapping rules to support automatic FHIR instances population from the heterogeneous clinical database or NLP output types from multiple NLP tools; (2) the creation of normalization rules to support non-standard data content transforming into standard FHIR representation; (3) the definition of a collection of NLP-specific FHIR extensions to enhance the FHIR model supportability for unstructured data; and (4) the construction of the FHIR-based type system used for improving interoperability among existing NLP tools and components. The design architecture supports extensibility and scalability as the FHIR-based type system covers all core clinical resources in the FHIR specification, which makes the NLP2FHIR pipeline modular. For instance, we can easily extend the architecture to produce a new data normalization pipeline profile using the FHIR DiagnosticReport resource to support the modeling of unstructured diagnostic reports (eg, pathology or radiology reports) in the future.

The NLP2FHIR pipeline provides a generic and scalable framework to support the FHIR modeling of unstructured EHR data. We have focused on the use of the core clinical resources Condition, Procedure, MedicationStatement (including Medication), and FamilyMemberHistory. We needed to handle those FHIR elements that were not covered by the NLP outputs through investigating: (1) whether the values of the elements could be retrieved using structured data (Table 3); and (2) whether new relationship detection algorithms should be developed for a specific element (eg, FamilyMemberHistory.relationship). We solicited a collection of such elements and developed corresponding FHIR extensions (Table 2) within the NLP2FHIR pipeline. We argue that community-based consensus development is a critical next step to broaden the applicability of the NLP2FHIR pipeline in the clinical informatics research community.

Meanwhile, we identified several barriers to EHR data modeling using FHIR. First, we noticed that while some of the NLP output types could be directly mapped to FHIR elements without semantic differences, in most other cases, there were semantic gaps between data models in the existing NLP systems and the FHIR specification. Second, the content normalization work in FHIR remains a

challenging task as it depends to a large extent on both the external terminology services and the FHIR internal value sets. Many of the elements in an FHIR resource are associated with a list of coded values (ie, a value set); some are in the form of a set of fixed values defined in the FHIR specification, while others are in the form of a list of concept codes defined in external terminologies or ontologies (eg, LOINC,<sup>21</sup> RxNorm,<sup>22</sup> or SNOMED CT<sup>23</sup>) If needed, a locally maintained dictionary and/or look-up table can even be used as a part of an FHIR profile. Currently, the FHIR code system and value set lists are under construction,<sup>24</sup> and integrating FHIR terminology services into our pipeline is critical for the future work. Fortunately, a number of community efforts have been initiated in developing open FHIR terminology services, including HAPI FHIR Terminology Loader for SNOMED CT and LOINC,<sup>25</sup> LOINC FHIR Terminology Server,<sup>26</sup> and Health Open Terminology FHIR Server.<sup>27</sup> Third, the mapping and content normalization rules are executed as part of transformation script within our NLP2FHIR pipelines. For future work, we plan to adopt formal models like the FHIR StructureMap resource to represent those structure mapping rules and the ConceptMap resource to represent the content normalization rules. This would enable an automated conversion process to be standardized by the FHIR specification.

## CONCLUSION

In this study, we developed and evaluated a standards-based clinical data normalization pipeline to model EHR data using the FHIR specification. We demonstrated that our NLP2FHIR pipeline is feasible for standardizing unstructured EHR data and integrating structured data into the model. The outcomes of this work provide standards-based tools of clinical data normalization that is indispensable for enabling portable EHR-driven phenotyping and large-scale data-driven analytics, as well as useful insights for future development of the FHIR specification on the handling of unstructured clinical data. With the standards-based FHIR modeling of both structured and unstructured EHR data, the NLP2FHIR pipeline would greatly benefit electronic health care data exchange,

utilization, and rapid generation of computational data for advancing clinical and translational research. We are actively working on improving the performance of the NLP2FHIR pipeline, developing new pipeline profiles with more FHIR clinical resource support, and applying the pipeline for EHR-driven cohort identification and data analytics. To accelerate community collaboration and tooling validation, the source code of our tooling and related resources are publicly available at the GitHub site: <https://github.com/BD2KOnFHIR/NLP2FHIR>.

## FUNDING

This work was supported by the National Institutes of Health (NIH) grant number U01 HG009450.

## AUTHOR CONTRIBUTIONS

N.H. and G.J. conceived the study design; N.H., A.W., and G.J. drafted the manuscript; A.W. and N.H. led the NLP2FHIR pipeline development and evaluation; A.W., F.S., S.S., C.W., N.H., and G.J. participated in the data analysis and result interpretation; H.L. and G.J. provided leadership for the project; all authors contributed expertise and edits. All authors read and approved the final version of the manuscript.

## ETHICS STATEMENT

This study has been reviewed and approved by the Mayo Clinic Institutional Review Board.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## ACKNOWLEDGMENTS

A preliminary work focusing on the data normalization profile for medication data was presented at the AMIA 2018 Informatics Summit.

## CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

## REFERENCES

- HL7 FHIR. 2019. <https://www.hl7.org/fhir/>. Accessed October 8, 2019.
- HL7 Argonaut Project. 2019. [http://argonautwiki.hl7.org/index.php?title=Main\\_Page](http://argonautwiki.hl7.org/index.php?title=Main_Page). Accessed October 8, 2019.
- Sync for Science Project. 2019. <http://syncfor.science/>. Accessed October 8, 2019.
- 21ST Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program Proposed Rule. 2019. <https://www.healthit.gov/sites/default/files/nprm/ONCCuresNPRMAPICertification.pdf>. Accessed October 8, 2019.
- FHIR Composition. 2019. <https://www.hl7.org/fhir/composition.html>. Accessed October 8, 2019.
- Savova GK, Tseytlin E, Finan S, et al. DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records. *Cancer Res* 2017; 77 (21): e115–e8.
- Hochheiser H, Castine M, Harris D, Savova G, Jacobson RS. An information model for computable cancer phenotypes. *BMC Med Inform Decis Mak* 2016; 16 (1): 121.
- Wu ST, Kaggal VC, Dligach D, et al. A common type system for clinical natural language processing. *J Biomed Semantics* 2013; 4 (1): 1.
- Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
- SHARPN Clinical Element Models. 2019. <http://informatics.mayo.edu/sharp/index.php/CEMS>. Accessed October 8, 2019.
- Chute CG, Pathak J, Savova GK, et al. The SHARPN project on secondary use of Electronic Medical Record data: progress, plans, and possibilities. *AMIA Annu Symp Proc* 2011; 2011: 248–56. PubMed PMID: 22195076; PMCID: PMC3243296.
- Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. *J Am Med Inform Assoc* 2014; 21 (5): 858–65.
- Wang Y, Wang L, Rastegar-Mojarad M, Liu S, Shen F, Liu H. Systematic analysis of free-text family history in electronic health record. *AMIA Jt Summits Transl Sci Proc* 2017; 2017: 104–13.
- Sohn S, Waghlikar KB, Li D, et al. Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification. *J Am Med Inform Assoc* 2013; 20 (5): 836–42.
- HAPI FHIR Java API. 2019. <https://hapifhir.io/>; Accessed October 8, 2019.
- Hong N, Wen A, Shen F, et al. Integrating structured and unstructured EHR data using an FHIR-based type system: a case study with medication data. *AMIA Jt Summits Transl Sci Proc* 2018; 2018: 74–83.
- NOTE\_NLP table. 2019. [http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:note\\_nlp](http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:note_nlp). Accessed October 8, 2019.
- FHIR Bundle. 2019. <https://www.hl7.org/fhir/bundle.html>. Accessed October 8, 2019.
- Hong N, Wen A, Mojarad MR, Sohn S, Liu H, Jiang G. Standardizing heterogeneous annotation corpora using HL7 FHIR for facilitating their reuse and integration in clinical NLP. *AMIA Annu Symp Proc* 2018; 2018: 574–83.
- Jha AK. Meaningful use of electronic health records: the road ahead. *JAMA* 2010; 304 (15): 1709–10.
- LOINC. 2019. <https://loinc.org/>; Accessed October 8, 2019.
- RxNorm. 2019. <https://www.nlm.nih.gov/research/umls/rxnorm/>. Accessed October 8, 2019.
- SNOMED CT. 2019. <http://www.snomed.org/snomed-ct/>. Accessed October 8, 2019.
- Walls RL, Athreya B, Cooper L, et al. Ontologies as integrative tools for plant science. *Am J Bot* 2012; 99 (8): 1263–75.
- Command Line Tool for HAPI FHIR. 2019. [https://hapifhir.io/doc\\_cli.html](https://hapifhir.io/doc_cli.html); Accessed October 8, 2019.
- LOINC FHIR Terminology Server. 2019. <https://loinc.org/fhir/>. Accessed October 8, 2019.
- Health Open Terminology FHIR Server. 2019. <https://ctsa.ncats.nih.gov/cd2h/health-open-terminology-fhir-server/>. Accessed October 8, 2019.